의사결정나무를 이용한 방문학습지사의 고객세분화에 관한 연구

서광규*, 오은주*, 한영규*, 심현정*
*상명대학교 산업정보시스템공학과
e-mail : kwangkyu@smu.ac.kr

A Study on Customer Segmentation of the Home Study Company using Decision Tree

Kwang-Kyu Seo*, Yeun-Joo Oh*, Young-Kyu Han*,
Hyun-Jeong Shim*
*Department of Industrial Information and Systems Engineering,
Sangmyung University

요 약

Due to keen competition among companies, companies have segmented customers and they are trying to offer specially targeted customer by means of the distinguished method. In accordance, data mining techniques are noted as the effective method that extracts useful information. This paper explores customer segmentation of the home study company using data mining. The purposes of this paper are especially competitor churn in the recent home study market, to understand the characteristics of the customer group who are expected churn in case competing companies do aggressive sales promotion. In addition, this paper aims to find the influential factors of their breakaway, and to prepare practical marketing strategy to keep the existing customers. The study of churn in the home study market is conducted and the model using decision tree to predict and select valuable customer. Finally, this paper presents how the results can be incorporated and measured as a part of an overall marketing campaign process.

1. 서론

현대의 많은 기업들은 이미 시장의 포화상태에서 경쟁을 하고 있고, 이러한 경쟁사회에서는 작은 정보하나라도 기업에게 큰 영향을 끼치게 된다. 대분분의 기업들은 무형적으로든 유형적으로든 방대한 양의 정보를 축적하고 있는데, 경쟁력을 확보하기 위해서는 축적된 데이터들 중에서 유용한 것들을 골라내고 이것을 분석하는 것이 중요하다. 데이터마이닝은 대용량의 데이터로부터 유용하게 활용될 수 있는 지식을 효과적으로 찾아내는 지식 탐사의 한 연구 분야로써, 수많은 형태의 방대한 데이터에서 그 각각의 목적에 도움이 될 수 있는 유용한 지식을 추출하는 것이다. 최근에는 기업 업무의 효율적 수행을 위해 데이터베이스를 이용하고, 그 결과를 활용하는 단계로부터 데이터 자체의 분석을 통해 행동패턴을 추출해내고 이 결과를 업무와 생산의 효율성 증대를 위해 이용하는 단계로 넘어서는 추세인데, 이런 추세에 맞추어 데이터마이닝 기술이 소개되었고, 이는 기업의 경쟁력 확보와 문제점 파악을 위한 기반 기술로 많은 발전이 이루어지고 있다 [1, 2, 6].

사교육적 차원에서 학원이나 과외 말고는 마땅히 교육 시스템이 없었던 초기의 방문 학습지 시장은 비교적 경쟁이 약한 시장이었다. 굵직한 몇 개의 기업만이 업계의 전부였고, 다양하지 못한 상품 등으로 고객들은 선택의 폭이 좁았다. 대한민국의 학부모들은 다른 아이들이 하는 것을 내 자식이 하지 않

으면 혹시나 뒤쳐질까 하는 불안심리를 갖고, 좋든 싫든 방문 학습지를 시키는 경우가 많았다.

그러나 근래에 들어서는 여러 기업이 이 업계에 뛰어들고 있고, 그만큼 다양한 상품과 구성으로 소비자들의 입맛을 사로잡고 있는 상황이 되면서 이들 기업간의 경쟁은 점점 심화되었다. 그리고 방문 학습지 시장의 가장 큰 타깃인 초등학생층은 한계가 있기 때문에 고객 쟁탈전 또한 심화되고 있다. 이러한 시점에서 다른 여러 업계와 마찬가지로 방문 학습지 업계 또한 신규고객의 유치보다는 코스트 효율이 높은 기존고객의 이탈을 방지하는 측면에 한층 무게중심을 두고 있다.

본 연구는 데이터마이닝을 통하여 G학습지사의 기존고객과 이탈고객의 성향을 분석하고, 이탈고객의 패턴을 추출하여 향후 마케팅 전략을 수립에 도움을 주고자 한다. 실제 자료 분석은 SPSS사의 SPSS 10.0 [3]과 동일회사 데이터마이닝 도구인 Clementine 7.0 [7]을 사용하였다.

2. 의사결정나무

데이터 마이닝 기법중 의사결정나무(Decision Tree) 분석은 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 분석과정이 나무구조에 의해서 표현되기 때문에 신경망(Neural Networks), 판별분석(Discriminant Analysis), 회기분석과 같은 방법들에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다.

본 연구에서는 의사결정나무 알고리즘의 대표적인 CART를 이용하여 방문학습지사의 고객세분화 연구를 수행하기로 한다.

CART(Classification and Regression Tree) 알고리즘은 의사결정나무를 형성하는데 있어서 가장 보편적인 알고리즘이다. CART는 지니(범주형 목표변수인 경우 적용) 또는 분산의 감소량(연속형 목표변수인 경우 적용)을 이용하여 이진 분리(binary split)를 수행하는 알고리즘이다. 지니지수(Gini Index)는 불순도(impurity)를 측정하는 하나의 지수이다. 임의의 한 개체가 목표변수의 i 번째 범주로부터 추출되었고, 그 개체를 목표변수의 j 번째 범주에 속한다고 오분류(misclassification)할 확률은 $P(i)P(j)$가 된다. 여기에서 $P(i)$는 각 마디에서 한 개체가 목표변수의 i번째 범주에 속할 확률이다. 이러한 오분류 확률을 모두 더하여

$$G=\sum_{j=1}^{c}\sum_{i\neq j}P(i)P(j)$$

을 얻을 수 있고 이는 위와 같은 분류 규칙하에서 오분류 확률의 추정치가 된다. 여기서 c 는 목표변수의 범주의 수를 말한다. 일반적으로 CART는 범주형 목표변수에 대해서는 지니지수를 분리기준으로 사용한다. 지니지수는 각 마디에서의 불순도 또는 다양도(diversity)를 재는 측도 중의 하나로써

$$G=\sum_{j=1}^{c}P(j)(1-P(j))=1-\sum_{j=1}^{c}P(j)^2=1-\sum_{j=1}^{c}(\frac{n_j}{n})^2$$

와 같이 표현될 수 있다. 여기에서 n은 그 마디에 포함되어 있는 관찰치수를 말하고, n_j는 목표변수의 i 번째 범주에 속하는 관찰치수를 말한다. 지니 지수는 n개의 원소 중에서 임의로 2개를 추출하였을 때 추출된 2개가 서로 다른 그룹에 속해 있을 확률을 의미하며 Simpson의 다양도 지수(diversity index)로도 알려져 있다. 목표변수의 범주가 2개인 경우에는 지니 지수는 다음과 같이 표현될 수 있다.

$$G=2P(1)P(2)=2(\frac{n_1}{n})(\frac{n_2}{n})$$

이는 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다. CART 알고리즘은 지니 지수를 가장 감소시켜 주는 예측변수와 그 변수의 최적분리를 자식마디로 선택하는데, 지니 계수의 감소량은 다음과 같이 계산된다.

$$\Delta G=G-\frac{n_L}{n}G_L-\frac{n_R}{n}G_R$$

여기서 n은 부모마디의 관측치 수를 말하고 n_R과 n_L는 각각 자식마디의 관측치 수를 의미한다. 즉, 자식마디로 분리되었을 때의 불순도가 가장 작도록 자식마디를 형성하는 것이다. 이는 다음과 같은 자식마디에서의 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)G_L+P(R)G_R=\frac{n_L}{n}G_L+\frac{n_R}{n}G_R$$

2. 의사결정나무 분석

본 논문에서는 G 방문학습지사의 학습지를 현재 구독하고 있는 고객들과 이미 이탈한 고객들의 설문지를 이용한 고객세분화와 고객이탈분석을 수행하였다. 유지고객의 만족도와 불만도 등 여러 항목들을

통해 패턴을 읽어내어 G 방문학습지사에서 어떠한 요인들이 고객에게 영향을 끼치고 있는지 또한 반면에 본 학습지사의 이탈고객의 만족도와 불만도 등의 패턴을 읽어내어 어떠한 요인들의 영향 때문에 이탈을 했는지 알아보는 것이 주요 관점이었다.

3.1. 분석자료 및 방법

분석에 사용된 자료는 고객들을 대상으로 실시한 설문지 400장이 이용되었다. 이중 유지고객 250명(62.5%), 이탈고객 150명(37.5%)이다. 목표변수는 이탈고객은 1(무)의 값을 주고 유지고객은 0(유)의 값을 갖는 변수로 설정하였다. 설명변수로는 설문지에 작성된 항목들로 구성되었다. 설명변수는 총 34개로 연속형 변수와 명목형 변수가 함께 사용되었다.

분석방법은 고객이탈 여부를 알려주는 목표변수(target variable)가 있으므로 관리학습(Supervised learning)방법, 의사결정나무 CART를 이용하여 분석하였다. 400개의 데이터 중에서 랜덤으로 추출한 결과 Test_data는 총 208개(유지고객 : 135명(64.90%), 이탈고객 : 73명(35.1%)), Train_data는 총 192개(유지고객 : 115명(59.89%), 이탈고객 : 77명(40.11%))로 분할된 데이터들의 분포는 전체데이터와 비슷하게 구성되어 있음을 알 수 있다.

3.2. CART를 이용한 고객이탈분석

분할과정 후 Test_data로 우선 CART를 학습시키는 과정을 거친다. 이탈여부, 변경시기 를 예측하는 CART 알고리즘을 이용한 모형구축을 실시하였고, 분석 흐름도는 그림 1과 같다.

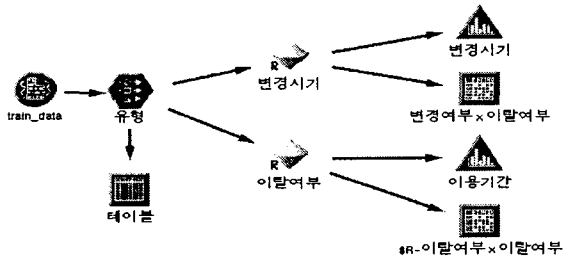

그림 1. CART 알고리즘 모형

우선 CART로 분석하기 전에 SPSS의 요인분석 결과에서는 가격, 방문횟수, 학습시간, 선생님의 조건이 요인으로 나왔다. CART에서는 고객 이탈여부에 영향을 미치는 중요한 변수로는 학습시간, 선생님의 조건, 가격, 학습지 내용, 방문횟수, 선택경로로 나타났다. CART의 요인들이 더 자세히 나왔지만 SPSS에서의 요인과 동일하다는 것을 알 수 있다.

여기서 CART는 순환적 분리를 수행하므로 특정 입력변수에 의해 나누어진 뒤에도 재차 동일한 변수에 의해 나누어지고 있음을 볼 수 있다. 괄호 안에의 두 숫자는 해당 노드 내에서 관찰치의 빈도와 특정 목표범주가 점유하는 비율을 나타낸다.

CART 알고리즘을 통하여 얻은 결과는 그림 2와 같다.

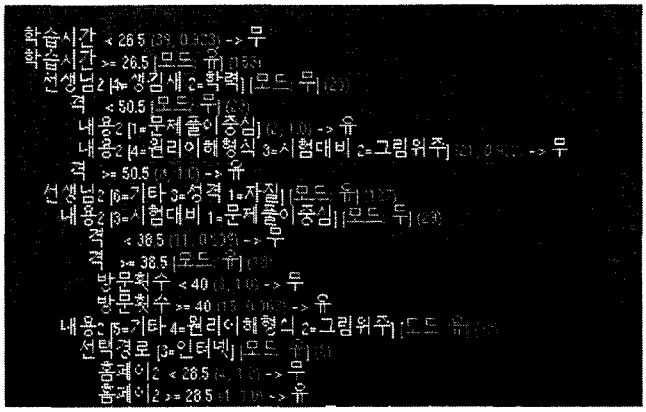

그림 2. CART 알고리즘을 통하여 얻은 결과

이탈가능성이 높은 집단은 학습시간을 가장 중요한 요소로 보며, 학습시간에 대한 만족도가 26.5점 미만인 고객이며, 전체 192명 중에 39명이 이러한 특성을 가진 고객이며 다른 이탈가능성 집단으로는 학습지 내용 중 원리이해형식을 중요하게 생각하는 고객 21명이 있으며 95.2%가 이탈하고 있다. 또 다른 이탈가능성이 높은 집단은 학습지 내용 중 시험과 관련한 내용들을 중요하게 생각하면서 동시에 가격에 대한 만족도가 38.5점 미만인 고객들이다. 고객 11명이 있으며 90.9%가 이탈하고 있다. 방문횟수의 만족도가 40점 미만인 고객 3명과 홈페이지를 이용할 의향에 대해서 28.5점미만인 점수를 준 4명은 100% 이탈하고 있다.

유지가능성이 높은 집단은 학습시간에 대한 만족도가 26.5점 이상이며, 문제풀이가 중심이 된 학습지 내용을 선호하는 고객 2명과 가격에 대한 만족도가 50.5점 이상인 고객 3명이 학습지를 구독하고 있다. 또 다른 집단으로는 선생님의 성격과 자질을 중요하게 생각하면서 현 방문횟수인 1주일에 한번 방문하

는 것에 대해 40점 넘게 만족하는 고객들 15명, 86.7%가 유지되고 있다. 또한 인터넷으로 학습지를 선택한 고객들 중 홈페이지를 사용할 의향에 대해서 28.5점 이상의 점수를 준 고객 4명은 100% 학습지를 구독하고 있다.

위에서 적용된 모형에 의한 결과를 오분류율을 중심으로 평가해 보면 그림 3과 같다.

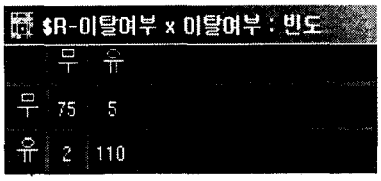

그림 3. 오분류율표

실제의 유지고객은 115명(59.89%)이고, 이탈고객은 77명(40.11%)이다. 예측된 유지고객은 112명(58.33%)이고, 이탈고객은 80명(41.67%)이다. 구현된 모형이 전체 192개의 관찰치 중 유지를 유지로 분류했거나, 이탈을 이탈로 분류한 경우는 75+110=185로, 전체관찰치중 96.35%를 제대로 분류했음을 알 수 있다. 따라서 오분류율은 1-0.9635=0.0365가 된다.

본 연구에 사용된 서비스 품질 측정도구가 현상을 일관성 있게 측정하고 있는가를 보기 위하여 신뢰성 검증을 실시하였다. 신뢰성이란 일관성, 예측가능성, 정확성 등의 의미를 함축하는 개념으로, 비교가능한 측정방법에 의해 대상을 측정하는 경우 결과가 비슷하게 되는 것을 말한다. 신뢰성의 의의는 어떤 조사결과에 대하여 이 조사결과가 부정확한 측정자료에 의해서 우연히 발견된 것이 아니라는 확신을 줄 수 있다.

4. 결론

본 연구에서는 현존하고 있는 G학습지사의 유지고객과 이탈고객을 중심으로 데이터를 얻어 데이터마이닝의 관리학습기법 중 CART를 적용해 고객이탈에 영향을 주는 요인과 이탈고객집단을 분류하는 방법을 살펴보았다.

고객이탈과 관련이 깊은 변수로는 학습시간, 선생님의 조건, 가격, 학습지 내용, 방문횟수, 선택경로 등으로 나타났다. 요인 중 방문횟수에 대한 만족도 점수가 40점 이상에서는 이탈율이 줄어들고 있다. 또한 5개월이 지나면 변경하는 횟수 즉, 이탈하는 고객들이 현저히 줄어들고 있었고, 이용기간은 20개월이 지나면서 이탈율이 늘어나고 있고 4~6개월 사이에 이탈율이 가장 최고조에 달하는 것을 알 수 있다.

이 모형을 통해서 고객들의 이탈과 유지를 결정짓는 시기는 5개월 내외라는 것을 파악했다. 회사는 고객과의 관계를 꾸준히 지속적으로 유지하려면 5개월 내외의 시기에 고객에 대한 관리를 더욱 철저히 함으로써 이탈을 방지해야한다. 그러기 위해서는 처음 신규고객을 유치한 후부터의 친절한 서비스와 고객 개개인의 정보를 신속히 파악하여 고객의 니즈를 만족시켜야 하고, 불만의 처리도 신속하고 정확하게 받아들여야만 할 것이다.

참고문헌

[1] 강현철 외, *데이터마이닝 - 방법론 및 활용*, 자유아카데미 2001.

[2] 김영아, *시장세분화를 위한 데이터마이닝 응용에 관한 연구*, 서강대학교 석사학위 논문, 2001.

[3] 노형진, *한글 SPSS 10.0에 의한 조사방법 및 통계분석*, 형설출판사, 2004.

[4] 이극노, "이동통신고객 분류를 위한 의사결정나무와 신경망 결합 알고리즘에 관한 연구.", 한국지능정보시스템학회논문지 9(1) (2003) : 139-155

[5] 이현정, *데이터마이닝을 이용한 보험회사 고객이탈분석에 관한 연구*, 중앙대학교 석사학위 논문, 2001

[6] 조혜정, *고객세분화를 위한 데이터마이닝 기법 비교*, 동아대학교 석사학위 논문, 2001.

[7] 허준 외, *Clementine 7 매뉴얼*, SPSS 아카데미, 2003.